Boosting Transferability in Vision-Language Attacks via Diversification along the Intersection Region of Adversarial Trajectory

Sensen Gao [1,3] ⋆, Xiaojun Jia [2*†], Xuhong Ren [2],
Ivor Tsang [2,3], and Qing Guo [3†]

[1] Nankai University, Tianjin, China
[2] Nanyang Technological University, Singapore
[3] CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore

Abstract. Vision-language pre-training (VLP) models exhibit remarkable capabilities in comprehending both images and text, yet they remain susceptible to multimodal adversarial examples (AEs). Strengthening attacks and uncovering vulnerabilities, especially common issues in VLP models (*e.g.*, high transferable AEs), can advance reliable and practical VLP models. A recent work (*i.e.*, Set-level guidance attack) indicates that augmenting image-text pairs to increase AE diversity along the optimization path enhances the transferability of adversarial examples significantly. However, this approach predominantly emphasizes diversity around the online adversarial examples (*i.e.*, AEs in the optimization period), leading to the risk of overfitting the victim model and affecting the transferability. In this study, we posit that the diversity of adversarial examples towards the clean input and online AEs are both pivotal for enhancing transferability across VLP models. Consequently, we propose using diversification along the intersection region of adversarial trajectory to expand the diversity of AEs. To fully leverage the interaction between modalities, we introduce text-guided adversarial example selection during optimization. Furthermore, to further mitigate the potential overfitting, we direct the adversarial text deviating from the last intersection region along the optimization path, rather than adversarial images as in existing methods. Extensive experiments affirm the effectiveness of our method in improving transferability across various VLP models and downstream vision-and-language tasks. Code is available at `https://github.com/SensenGao/VLPTransferAttack`.

Keywords: Vision-Language Attack · Adversarial Transferability · Diversification · Intersection Region of Adversarial Trajectory

1 Introduction

Vision-language pre-training (VLP) models utilize multimodal learning, leveraging large-scale image-text pairs to bridge the gap between visual and lan-

⋆ co-first authors; † co-corresponding authors

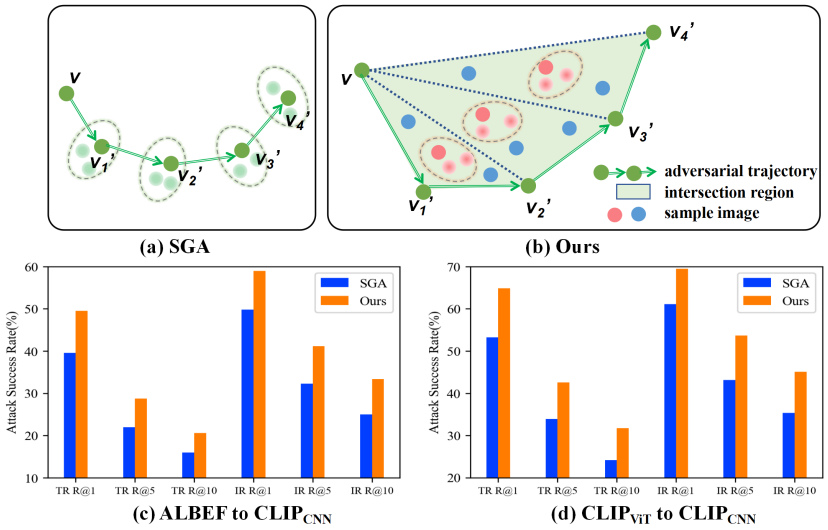

Fig. 1: Our method vs. set-level guided attack (SGA) [36]. (a) shows the main idea of SGA, *i.e.*, conducting augmentation around the online adversarial examples. (b) shows the main idea of our method, that is, we perform augmentation in the intersection region of adversarial trajectory. The red and blue dots both depict images sampled from the intersection region, with red dots indicating the best samples selected using the text-guided adversarial example selection strategy. The surrounding light red dots represent applying the same resizing data augmentation to the best samples as utilized in SGA. (c) and (d) compare the transferability of our method and SGA by using the adversarial examples of ALBEF [30] and CLIP_{ViT} to attack CLIP_{CNN}, respectively.

guage understanding. These models showcase revolutionary performance across various downstream Vision-and-Language tasks, including image-text retrieval, image captioning, visual grounding, and visual entailment, as demonstrated in [20, 27, 28, 44]. Despite their success, recent research underscores the significant vulnerability of VLP models, particularly when confronted with multimodal adversarial examples [7,12,16,17,36,37,57]. Uncovering the vulnerabilities, especially common issues, can drive further research aimed at building more reliable and practical VLP models.

Current research predominantly concentrates on attacking VLP models via a white-box setting, where the model's structural information can be exploited. However, exploring the transferability of multimodal adversarial examples is pivotal, especially given the limited access to detailed model structures in real-world scenarios. There have been some efforts made to enhance the transferability of attacks on VLP models by introducing input diversity, as seen in the work SGA [36]. While they have achieved some effectiveness, how to enhance the transferability of multimodal adversarial examples is still an open question.

In this paper, we undertake a comprehensive examination of the factors contributing to the limited transferability of the cutting-edge multimodal attack

method, *i.e.*, SGA [36]. As shown in Figure 1 (a), throughout the iterative generation of subsequent adversarial images, SGA conducts data augmentation around the online adversarial image (*i.e.*, adversarial examples generated during optimization). This strategy enhances the diversity of adversarial examples along the optimization path, leading to a certain improvement in transferability. However, such an approach still carries the potential risk of overfitting to the victim model with the high reliance on the examples along the adversarial trajectory (See Figure 1 (b)), which leads to low attack success rates when we migrate the adversarial example to other VLP models (See Figure 1 (c) and (d)). To mitigate this overfitting risk, one potential solution is to further enhance the diversity of augmented adversarial examples in a judicious manner.

Building upon the aforementioned analysis, SGA overfits local adversarial examples, while the clean image is the only accessible example that is far from local adversarial examples. Therefore, we embark on a pioneering endeavor to enhance the transferability of multimodal adversarial attacks by considering the diversity of adversarial examples (AEs) around clean inputs and online AEs throughout the optimization process. To achieve this, we consider the intersection region of adversarial trajectory, which encompasses the original image, the adversarial image from the previous step, and the current adversarial image during the iterative attack process (depicted in Figure 1 (b)). This innovative approach aims to circumvent overfitting by strategic sampling within this region, thereby avoiding an undue focus on adversarial example diversity solely around adversarial images. After obtaining multiple samples, we calculate gradients for each to determine perturbation directions away from the text. Subsequently, we individually incorporate these perturbations into the current adversarial image and select the one that deviates the most from the text.

Additionally, in the text modality, SGA only considers deviating the text from the last adversarial image in the optimization period, but the adversarial image is solely generated by the surrogate model, still posing the risk of overfitting the surrogate model. For this reason, we propose to have the text deviate simultaneously from the last intersection region along the optimization path.

Our proposed method is evaluated on two widely recognized multimodal datasets, Flickr30K [41] and MSCOCO [35]. We conduct experiments on three vision-and-language downstream tasks (*i.e.*, image-text retrieval (ITR), visual grounding(VG), and image captioning (IC)), and all results indicate the high effectiveness of our method in generating more transferable multimodal adversarial examples. Moreover, when adversarial examples generated from image-text retrieval are transferred to other vision-and-language downstream tasks (*i.e.*, VG and IC), there is a substantial improvement in attack performance.

Our main contributions can be summarized as follows:

- We propose using the intersection region of adversarial trajectory to expand the diversity of adversarial examples during optimization, based on which we develop a high-transferability attack against VLP models.

- We extend the generation of adversarial text to deviate from the last intersection region along the optimization path, aiming to reduce overfitting the surrogate model, thereby achieving enhanced transferability.
- Extensive experiments robustly demonstrate the efficacy of our proposed method in elevating the transferability of multimodal adversarial examples across diverse models and three downstream tasks.

2 Related Work

2.1 Vision-Language Pre-training Models

VLP models leverage multimodal learning from extensive image-text pairs to improve the performance of various Vision-and-Language (V+L) tasks [29]. Early VLP models predominantly depend on pre-trained object detectors for acquiring multimodal representations [6, 32, 45, 49, 59]. Recently, with the advent of end-to-end image encoders like the Vision Transformer (ViT) [10, 46, 56] offering faster inference speeds, some work propose to use them as substitution for computationally expensive object detectors [11, 29, 30, 48, 53].

There are two popular approaches for VLP models in learning vision-language representations: the fused architecture and the aligned architecture. Fused VLP models (*e.g.*, ALBEF [30], TCL [53]), initially employ two separate unimodal encoders to learn features for text and images. Subsequently, a multimodal encoder is utilized to fuse the embeddings of text and images. In contrast, aligned VLP models, exemplified by CLIP [42], focus on aligning the feature spaces of distinct unimodal encoders and benefit downstream tasks significantly [1]. This paper concentrates on evaluating our proposed method using multiple popular fused and aligned VLP models.

2.2 Downstream Vision-and-Language Tasks

Image-Text Retrieval (ITR) involves retrieving pertinent information, textual or visual, in response to queries from another modality [5, 8, 51, 60]. This undertaking usually encompasses two sub-tasks: image-to-text retrieval (retrieving text based on an image query) and text-to-image retrieval (retrieving images given a text query).

In aligned VLP models, both the Text Retrieval (TR) and Image Retrieval (IR) tasks leverage ranking results determined by the similarity between text and image embeddings. However, in fused VLP models, where internal embedding spaces lack alignment across unimodal encoders, the similarity scores between image and text modalities are computed for all image-text pairs to retrieve the Top-N candidates. Subsequently, these Top-N candidates serve as input for the multimodal encoder, which computes the image-text matching score to establish the final ranking.

Visual Grounding (VG) refers to the task of localizing the region within a visual scene with corresponding entities or concepts in natural language. Among

Table 1: Attack Success Rate (%) of SGA with and without image augmentation. The SGA w.o. Aug doesn't consider image augmentation. We use ALBEF to generate multimodal adversarial examples on the ITR task to evaluate transferability.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	SGA w.o. Aug	**99.9**	99.95	70.07	71.67	30.55	39.88	32.31	42.54
	SGA w. Aug	**99.9**	**99.98**	**87.88**	**88.05**	**36.69**	**46.78**	**39.59**	**49.78**

VLP models, ALBEF expands Grad-CAM [43] and utilizes the acquired attention map to rank the detected proposals [54].

Image Captioning (IC) is to generate a textual description that logically describes or implies the content of a given visual input, typically involving the creation of captions for images. The evaluation of Image Captioning models often employs metrics such as BLEU [39], METEOR [4], ROUGE [33], CIDEr [47] and SPICE [2], which serve to assess the quality and relevance of the generated captions in comparison to reference captions.

2.3 Transferability of Adversarial Examples

Adversarial attacks [13–15, 18, 25] are typically categorized as white-box and black-box attacks. In a white-box setting [23, 26], the attacker has full access to the model, whereas black-box attacks [3, 40], more realistic in practical applications, occur when information about the model is limited. In the realm of image attacks [21,22,24], prevalent methods for crafting transferable adversarial examples often leverage data augmentation techniques (*e.g.*, DIM [52], TIM [9], SIM [34], ADMIX [50], PAM [58]). Zhang *et al.* [57] introduced a white-box attack targeting popular VLP models for downstream tasks in the multimodal domain. Building upon this work, Lu *et al.* [36] proposed SGA, considering the diversity of adversarial examples by expanding single image-text pairs to sets of images and texts to conduct black-box attacks on VLP models.

However, SGA primarily emphasizes diversity in the vicinity of adversarial examples during the optimization process, potentially increasing the risk of overfitting the victim model and impacting transferability. Therefore, our primary focus in this study is to further enhance the diversity of adversarial examples in a thoughtful manner and avoid an undue focus on adversarial example diversity solely around adversarial images.

3 Methodology

3.1 Background and Motivation

Adversarial attacks on VLP models involve inducing a mismatch between adversarial images and corresponding adversarial text while adhering to specified

constraints on image and text perturbations. Here, (v, t) denotes an original image-text pair from a multimodal dataset, with v' representing an adversarial image and t' denoting adversarial text. The allowable perturbations are restricted within the ranges $B[v, \xi_v]$ for images and $B[t, \xi_t]$ for text. The image and text encoders of the multimodal model are denoted as F_I and F_T, respectively. To generate valid multimodal adversarial examples, the objective is to maximize the loss function J specific to VLP models:

$$\begin{cases} \max J\left(F_I(v'), F_T(t')\right) \\ s.t. v' \in B[v, \xi_v], t' \in B[t, \xi_t]. \end{cases} \tag{1}$$

The state-of-the-art approach for exploring the transferability of multimodal adversarial examples (*i.e.*, SGA [36]) involves augmenting image-text pairs to enhance the diversity of adversarial examples along the optimization path. Specifically, during the iterative generation of adversarial images, let v'_i represent the adversarial image generated at the i-th step. In the subsequent step $(i+1)$, SGA initiates the process by applying a resizing operation for data augmentation to v'_i, resulting in $V'_i = \{v'_{i1}, v'_{i2}, ..., v'_{iM}\}$ (See Figure 1 (a)). The iterative formula can be expressed as follows:

$$v'_{i+1} = v'_i + \alpha \cdot sign\left(\frac{\nabla_v \sum_{j=1}^{M} J(F_I(v'_{ij}), F_T(t))}{\|\nabla_v \sum_{j=1}^{M} J(F_I(v'_{ij}), F_T(t))\|}\right). \tag{2}$$

To further examine the impact of image augmentation along the optimization path in the SGA method, we utilize ALBEF as a surrogate model to generate multimodal adversarial examples. These examples are then employed to target VLP models such as TCL and CLIP, assessing the transferability of the attacks. Detailed results are presented in Table 1. Our observations reveal that SGA enhances the transferability of adversarial attacks, showing an increase ranging from 6.14% to 17.81%. However, it is noteworthy that the success rate of attacks on the target models remains notably lower than that on the source model. This discrepancy is primarily attributed to the fact that SGA predominantly emphasizes diversity around AE v'_i during the optimization period, without adequately considering the diversity of adversarial examples toward the clean image, leading to the risk of overfitting the victim model and affecting the transferability.

For this purpose, we propose to consider diversification along the intersection region of adversarial trajectory, which encompasses the original image v, the adversarial image from the previous step v'_{i-1}, and the current adversarial image v'_i during the iterative attack process. This region is established to sample images within it to broaden the diversity of adversarial examples (See Figure 1 (b)). Moreover, to fully leverage the interplay between modalities, we aim for perturbations guided by textual information that induce v'_i to deviate significantly from the associated text t. Additionally, in the text modality, our objective is to identify adversarial perturbations that simultaneously deviate from the intersection region rather than only adversarial images, thereby reducing overfitting the surrogate model and enhancing the effectiveness of black-box attacks.

3.2 Diversification along the Intersection Region

As outlined in Section 3.1, we enhance the diversity of adversarial examples by introducing diversification along the intersection region. Specifically, at the ith iteration during optimization, we have the v'_i, v'_{i-1}, and the clean v, and these variables form a triangle region denoted as $\triangle v v'_{i-1} v'_i$, *i.e.*, the intersection region of adversarial trajectory in Figure 1. Then, we initially sample multiple instances within the region $\triangle v v'_{i-1} v'_i$, representing the set of samples as $e = \{e_1, e_2, ..., e_N\}$. Each sample can be expressed as $e_k = \beta \cdot v + \gamma \cdot v'_{i-1} + \eta \cdot v'_i$, where $\beta + \gamma + \eta = 1.0$. Consequently, we can compute the gradient perturbation for each sample. For the k-th sample, denoted as e_k, its gradient perturbation p_k is calculated as follows. In this way, we can get a perturbation set $P = \{p_1, p_2, ..., p_N\}$ by

$$p_k = \alpha \cdot sign(\frac{\nabla_e J(F_I(e_k), F_T(t))}{\|\nabla_e J(F_I(e_k), F_T(t))\|}). \tag{3}$$

3.3 Text-guided Augmentation Selection

In Section 3.2, a diverse perturbation set P is derived from the intersection region. To harness the full potential of modality interactions, we introduce text-guided augmentation selection to obtain the optimal sample. Specifically, we individually incorporate each element from the perturbation set P into the adversarial image v'_i. The selection process aims to identify the sample that maximally distances v'_i from t. This procedure can be represented as:

$$m = \arg\max_{p_m \in P} J(F_I(v'_i + p_m), F_T(t)). \tag{4}$$

At this juncture, e_m represents the selected sample. We employ SGA as our baseline and incorporate the image augmentation methods considered along its optimization path. The chosen optimal sample e_m is resized and expanded into the set $E_m = \{e_{m1}, e_{m2}, ..., e_{mM}\}$. Subsequently, we utilize the expanded set E_m to generate the final adversarial perturbation, yielding v'_{i+1}:

$$v'_{i+1} = v'_i + \alpha \cdot sign(\frac{\nabla_e \sum_{j=1}^{M} J(F_I(e_{mj}), F_T(t))}{\|\nabla_e \sum_{j=1}^{M} J(F_I(e_{mj}, F_T(t))\|}). \tag{5}$$

3.4 Adversarial Text deviating from the Intersection Region

In the text modality, SGA only considers deviating adversarial text from the ultimate adversarial image generated during the iterative optimization process. If there are a total of T iterations, The generation of t' only considers deviating from the adversarial image v'_T. The adversarial image v'_T is exclusively created by the surrogate model, thereby still presenting the risk of overfitting to the surrogate model. However, during the optimization process of the adversarial image, the clean image is entirely independent of the surrogate model. For this

reason, we propose to have the text deviate simultaneously from the last intersection region along the optimization path. Specifically, the adversarial text deviates from the triangle region constituted by v, v'_{T-1} and v'_T.

$$t' = \underset{t' \in B[t,\epsilon_t]}{\arg\max}(\lambda \cdot J(F_I(v), F_T(t')) + \mu \cdot J(F_I(v'_T), F_T(t'))) + \nu \cdot J(F_I(v'_{T-1}), F_T(t'))). \tag{6}$$

We also set adjustable scaling factors, among which $\lambda + \mu + \nu = 1.0$.

3.5 Implementation Details

In the specific process of our attack, we employ an iterative approach. In each iteration, we sample within the intersection region of adversarial trajectory, guided by textual information, to select a sample that maximally deviates the current adversarial image v'_i from the text t. Subsequently, we subject this sample to image augmentation processing, calculate gradients to determine the perturbation direction, and overlay it onto the current adversarial image, resulting in v'_{i+1}. Through multiple iterative steps, we obtain an adversarial image v'_T. For the text modality, in contrast to previous methods that solely focus on deviating from the adversarial image v'_T, our goal is to derive an adversarial text t' that simultaneously deviates from the last intersection region $\triangle v v'_{T-1} v'_T$ along the optimization path.

4 Experiments

In this section, we present experimental evidence demonstrating the enhanced transferability of multimodal examples generated from our proposed method across VLP models and various Vision-and-Language tasks. First, in Section 4.1, we introduce the experimental settings, including the popular image-text pair datasets and VLP models we use, as well as restrictions for adversarial attacks. Subsequently, the process of searching for optimal parameters is shown in Section 4.2. After that, we evaluate cross-model transferability in the context of the image-text retrieval task, as detailed in Section 4.3. Following this, in Section 4.4, we extend our investigation to the transfer of multimodal adversarial examples generated within the image-text retrieval task to other tasks, aiming to gauge cross-task transferability. Lastly, Section 4.5 outlines ablation studies.

4.1 Setups

VLP Models. In our transferability evaluation experiments across various VLP models, we explore two typical architectures: fused and aligned VLP models. We select CLIP [42] for the aligned VLP model. CLIP offers a choice between two distinct image encoders, namely ViT-B/16 [10] and ResNet-101 [19], denoted as CLIP_{ViT} and CLIP_{CNN}, respectively. In the case of fused VLP models, we opt for ALBEF [30] and TCL [53]. ALBEF uses a 12-layer ViT-B/16 image encoder

and two 6-layer transformers for text and multimodal encoding. TCL shares this architecture but has different pre-training objectives.

Datasets. In this study, we leverage two widely recognized multimodal datasets, namely Flickr30K [41] and MSCOCO [35], for evaluating the image-text retrieval task. The Flickr30K dataset comprises 31,783 images, each accompanied by five captions for annotation. Similarly, the MSCOCO dataset consists of 123,287 images, and approximately five captions are provided for each image.

Additionally, we employ the RefCOCO+ [55] dataset to assess the Visual Grounding task. RefCOCO+ is a dataset containing 141,564 referring expressions for 50,000 objects within 19,992 MSCOCO images. This dataset serves the purpose of evaluating grounding models by focusing on the localization of objects described through natural language. For another Vision-and-Language task, Image Captioning, we leverage the MSCOCO dataset as well.

Adversarial Attack Settings. In our study, we adopt adversarial attack settings of SGA [36] to ensure a fair comparison. Specifically, we leverage BERT-Attack [31] to craft adversarial texts. The perturbation bound ξ_t is set as 1 and length of word list $W = 10$. PGD [38] is employed to get adversarial images and the perturbation bound, denoted as ξ_v, is set as $8/255$. Additionally, iteration steps T is set as 10 and each step size $\alpha = 2/255$. Furthermore, when randomly sampling from the intersection region of adversarial trajectory, we set the number of samples to 5. When generating adversarial texts, we set the three parameters λ, μ, ν in Equation 6 to 0.6, 0.2, and 0.2 respectively. The values chosen for these adjustable parameters can be found in Section 4.2.

Evaluation Metrics. The key metric for adversarial transferability is the Attack Success Rate (ASR), which measures the percentage of successful attacks among all generated adversarial examples. A higher ASR indicates more effective and transferable attacks.

4.2 Optimal Parameters

In our proposed method, the number of samples N taken from the intersection region of adversarial trajectory and scaling factors in Equation 6 are adjustable. We conduct specific experiments to explore optimal parameter settings and examine their influence on the efficacy of our approach. To be more specific, we utilize ALBEF for generating multimodal adversarial examples on the Flickr30K dataset and assess the transferability on the other three VLP models.

Number of Samples taken from Intersection Region of Adversarial trajectory N. The bottom of Table 2 illustrates when the sample size reaches 5, transfer effects are observed for both the Image Retrieval task and the Text Retrieval task. As the sample size continues to increase, the transferability only fluctuates and does not exhibit further improvement. Taking into account both transfer effects and computational costs, a sample size of 5 is identified as the optimal configuration.

Scaling factors λ, μ, ν in Adversarial Text Generation. In Equation 6, λ represents the weight of clean images, while μ and ν represent the weights of adversarial images. Therefore, we stipulate that λ cannot be zero, and μ and ν

Table 2: Optimal Parameters: Attack Success Rate(%) on different settings, **Top** for different values of λ, μ, ν and **Bottom** for different numbers of samples N.

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	$[\lambda, \mu, \nu] = [\mathbf{0.2, 0.0, 0.8}]$	**99.9**	99.93	89.67	90.5	42.21	52.0	46.23	55.44
	$[\lambda, \mu, \nu] = [\mathbf{0.2, 0.2, 0.6}]$	**99.9**	99.93	90.41	90.43	41.96	51.87	46.49	55.3
	$[\lambda, \mu, \nu] = [\mathbf{0.2, 0.4, 0.4}]$	**99.9**	99.93	90.31	90.43	41.96	51.87	45.34	54.82
	$[\lambda, \mu, \nu] = [\mathbf{0.2, 0.6, 0.2}]$	**99.9**	**99.95**	90.52	90.57	42.21	51.84	45.08	54.92
	$[\lambda, \mu, \nu] = [\mathbf{0.2, 0.8, 0.0}]$	**99.9**	99.93	90.31	90.57	41.72	51.74	46.1	54.68
	$[\lambda, \mu, \nu] = [\mathbf{0.4, 0.0, 0.6}]$	**99.9**	99.93	91.25	90.88	45.52	55.32	48.91	57.63
	$[\lambda, \mu, \nu] = [\mathbf{0.4, 0.2, 0.4}]$	**99.9**	99.93	91.25	90.83	45.03	55.22	48.28	57.77
	$[\lambda, \mu, \nu] = [\mathbf{0.4, 0.4, 0.2}]$	**99.9**	99.93	91.15	90.71	44.91	55.28	48.15	57.87
	$[\lambda, \mu, \nu] = [\mathbf{0.4, 0.6, 0.0}]$	**99.9**	99.93	91.15	90.88	44.91	54.77	49.04	57.56
	$[\lambda, \mu, \nu] = [\mathbf{0.6, 0.0, 0.4}]$	**99.9**	99.93	91.46	90.95	**46.38**	56.38	49.04	**59.11**
	$[\lambda, \mu, \nu] = [\mathbf{0.6, 0.2, 0.2}]$	**99.9**	99.93	**91.57**	**91.17**	46.26	**56.8**	**49.55**	59.01
	$[\lambda, \mu, \nu] = [\mathbf{0.6, 0.4, 0.0}]$	**99.9**	99.93	90.94	90.98	46.01	56.72	**49.55**	58.87
	$[\lambda, \mu, \nu] = [\mathbf{0.8, 0.0, 0.2}]$	**99.9**	99.93	90.73	90.98	46.13	56.71	49.46	58.74
	$[\lambda, \mu, \nu] = [\mathbf{0.8, 0.2, 0.0}]$	**99.9**	99.93	90.31	90.95	45.77	56.65	49.34	58.87
ALBEF	$\boldsymbol{N = 3}$	**99.9**	**99.95**	90.62	90.79	45.64	56.54	48.83	58.73
	$\boldsymbol{N = 4}$	99.79	99.91	91.36	**91.17**	45.83	56.78	50.45	59.01
	$\boldsymbol{N = 5}$	**99.9**	99.93	**91.57**	**91.17**	**46.26**	56.8	49.55	59.01
	$\boldsymbol{N = 6}$	**99.9**	99.93	90.94	90.38	45.79	**56.96**	**50.7**	58.52
	$\boldsymbol{N = 7}$	**99.9**	99.91	89.88	90.95	45.4	56.35	**50.7**	**59.07**

can have at most one zero value. Initially, as the value of λ gradually increases, indicating the gradual introduction of clean images, the transferability of multimodal adversarial examples increases accordingly. However, when the value of λ becomes too large, it also leads to a disproportionately low proportion of adversarial images, resulting in a decrease in adversarial transferability. According to the experiments, the optimal parameters we select are $[0.6, 0.2, 0.2]$.

4.3 Cross-Model Transferability

As outlined in Section 4.1, our experimental design focuses on assessing the transferability of adversarial examples across two widely adopted VLP model architectures: fused and aligned. There are four VLP models selected, namely: ALBEF, TCL, CLIP_{ViT}, CLIP_{CNN}. The selected downstream V+L task for evaluation is image-text retrieval. Our approach involves employing one of four models to generate multimodal adversarial examples within the specified parameters of our proposed method. Subsequently, we validate the effectiveness of the generated adversarial examples through a comprehensive set of experiments, encompassing both self-attacks and attacks on the other three models. This evaluation encompasses one white-box attack and three distinct black-box attacks.

We adopt the methodology proposed by SGA [36] as our baseline, Therefore, the effectiveness of our method is compared against it. Moreover, we also present the effectiveness and transferability results of various other attack methods on VLP models. Table 3 provides a comprehensive comparison of these methods on the dataset Flickr30K, More experiments on the MSCOCO dataset are provided in the Appendix. In the comparison, PGD [38] is an image-only attack, while Bert-Attack [31] focuses solely on text-based attacks. Sep-Attack involves perturbing text and image separately. Furthermore, Co-Attack takes into account cross-modal interactions, generating an adversarial example for one modality

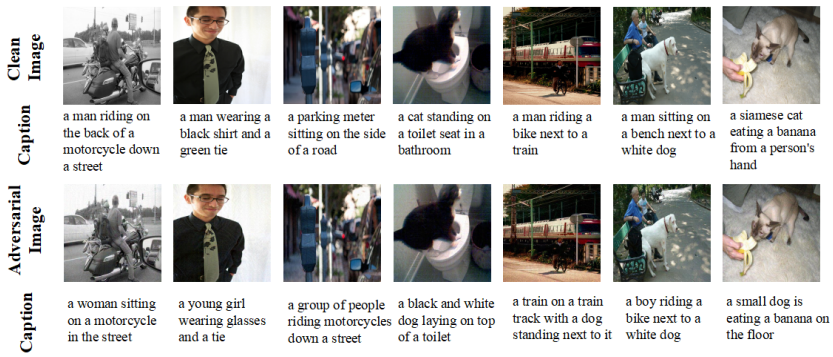

Fig. 2: Visualization on Image Captioning. We use the ALBEF model, pre-trained on Image Text Retrieval(ITR) task, to generate adversarial images on the MSCOCO dataset and use the BLIP [29] model for Image Captioning on both clean images and adversarial images, respectively.

under the guidance of the other modality. SGA, our baseline, expands a single image-text pair into a set of images and a set of texts to enhance diversity.

First, we compare the performance of various methods under white-box attacks, wherein we can leverage the model architecture and interact with the model. It is evident that our method, along with SGA, performs the best in the four white-box attack experiments compared to other methods. Whether TR or IR, the attack success rate at the top-1 rank (R@1) consistently exceeds 99.9%. Given the already high white-box attack success rate achieved by the SGA method, there is limited room for improvement in our approach within this context. Subsequently, we shift our focus to elucidating the enhancements our method brings to transferability, specifically in the realm of black-box attack performance. We delineate this exploration into two segments, contingent on whether the source model and the target model share the same architecture.

Cross-Model Transferability in Same Architecture. ALBEF and TCL both utilize a similar model architecture but differ in their pre-training objectives while maintaining a common fundamental model structure. Therefore, when ALBEF and TCL are employed as target models for each other, the success rate of attacks using multimodal adversarial examples is remarkably high. Notably, the black-box attack success rate of IR reaches 95.58% when TCL is the target model in our proposed method. Additionally, the transferability of SGA can reach approximately 90%, but the room for improvement is very limited, our method has improved compared to SGA, ranging from 2.71% to 3.69%. In contrast, when considering CLIP_{ViT} and CLIP_{CNN}—both being aligned VLP models—their image encoders vary significantly, with one utilizing the Vision Transformer and the other employing ResNet-101. Given the significant structural differences between traditional CNNs and Transformers, existing methods show low cross-model transferability, highlighting room for improvement. Our method outperforms SGA by about 7.67% to 11.62%.

Table 3: Comparison with state-of-the-art methods on image-text retrieval. The source column shows VLP models we use to generate multimodal adversarial examples. The gray area represents adversarial attacks under a white-box setting, the rest are black-box attacks. For both Image Retrieval and Text Retrieval, we provide R@1 attack success rate(%).

Source	Attack	ALBEF		TCL		CLIP_{ViT}		CLIP_{CNN}	
		TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1	TR R@1	IR R@1
ALBEF	PGD	93.74	94.43	24.03	27.9	10.67	15.82	14.05	19.11
	BERT-Attack	11.57	27.46	12.64	28.07	29.33	43.17	32.69	46.11
	Sep-Attack	95.72	96.14	39.3	51.79	34.11	45.72	35.76	47.92
	Co-Attack	97.08	98.36	39.52	51.24	29.82	38.92	31.29	41.99
	SGA	**99.9**	**99.98**	87.88	88.05	36.69	46.78	39.59	49.78
	Ours	**99.9**	99.93	**91.57**	**91.17**	**46.26**	**56.8**	**49.55**	**59.01**
TCL	PGD	35.77	41.67	99.37	99.33	10.18	16.3	14.81	21.1
	BERT-Attack	11.89	26.82	14.54	29.17	29.69	44.49	33.46	46.07
	Sep-Attack	52.45	61.44	99.58	99.45	37.06	45.81	37.42	49.91
	Co-Attack	49.84	60.36	91.68	95.48	32.64	42.69	32.06	47.82
	SGA	92.49	92.77	**100.0**	**100.0**	36.81	46.97	41.89	51.53
	Ours	**95.2**	**95.58**	**100.0**	99.98	**47.24**	**57.28**	**52.23**	**62.23**
$\textbf{CLIP}_{\text{ViT}}$	PGD	3.13	6.48	4.43	8.83	69.33	84.79	13.03	17.43
	BERT-Attack	9.59	22.64	11.80	25.07	28.34	39.08	30.40	37.43
	Sep-Attack	7.61	20.58	10.12	20.74	76.93	87.44	29.89	38.32
	Co-Attack	8.55	20.18	10.01	21.29	78.53	87.5	29.5	38.49
	SGA	22.42	34.59	25.08	36.45	**100.0**	**100.0**	53.26	61.1
	Ours	**27.84**	**42.84**	**27.82**	**44.6**	**100.0**	**100.0**	**64.88**	**69.5**
$\textbf{CLIP}_{\text{CNN}}$	PGD	2.29	6.15	4.53	8.88	5.4	12.08	89.78	93.04
	BERT-Attack	8.86	23.27	12.33	25.48	27.12	37.44	30.40	40.10
	Sep-Attack	9.38	22.99	11.28	25.45	26.13	39.24	93.61	95.3
	Co-Attack	10.53	23.62	12.54	26.05	27.24	40.62	95.91	96.5
	SGA	15.64	28.6	18.02	33.07	39.02	51.45	**99.87**	**99.9**
	Ours	**19.5**	**34.59**	**21.6**	**37.88**	**48.47**	**59.12**	**99.87**	**99.9**

Cross-Model Transferability in Different Architectures. When fused VLP models (*i.e.*, ALBEF and TCL) are used for generating adversarial examples, our method achieves a significant improvement compared to existing methods, outperforming the current state-of-the-art (SOTA) methods by 9.23% to 10.7%. When the fused VLP models are targeted for attack, our method still outperforms all existing methods when generating adversarial examples against the aligned VLP models (*i.e.*, CLIP). In this scenario, the transferability of all methods is relatively low. Compared to our method's baseline SGA, we attain an improvement ranging from 2.74% to 8.25%.

4.4 Cross-Task Transferability

We not only assess the transferability of multimodal adversarial examples generated by our proposed method across different VLP models but also conduct experiments to evaluate its effectiveness in transferring across diverse V+L tasks. Specifically, we craft adversarial examples for the Image-Text Retrieval (ITR) task and evaluate them on Visual Grounding (VG) and Image Captioning (IC) tasks. As evident from Table 4 and visual results in Figure 2 and 3, the adversarial examples generated for ITR demonstrate transferability, successfully impacting both VG and IC tasks. This highlights the efficacy of cross-task trans-

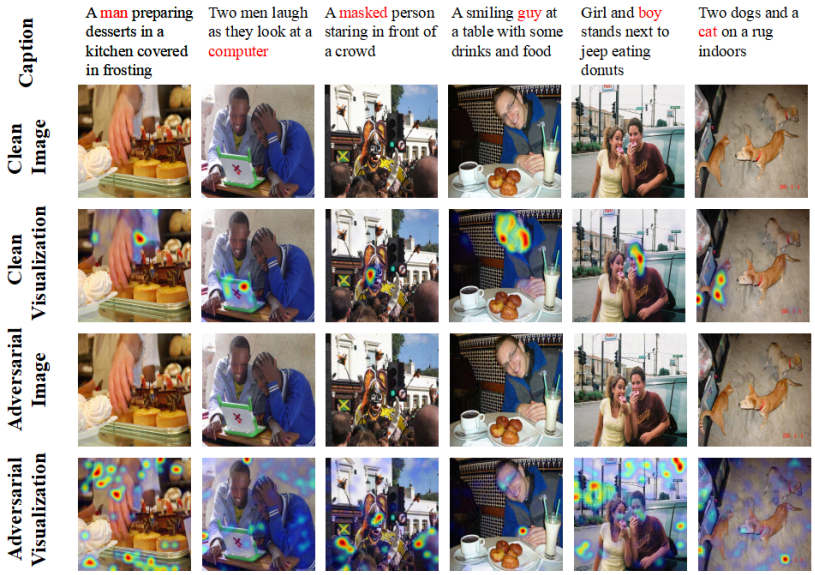

Fig. 3: Visualization on Visual Grounding. We use the ALBEF model, pre-trained on the ITR task, to generate adversarial images on the RefCOCO+ dataset and use the same model, pre-trained on Visual Grouding(VG) task, to localize the regions corresponding to red words on both clean images and adversarial images, respectively.

ferability in our proposed method. Furthermore, our transferability consistently outperforms that of SGA.

4.5 Ablation Study

Our proposed method builds upon the SGA [36] as a baseline, introducing two key improvements. Firstly, we utilize diversification along the intersection region of adversarial trajectory to expand the diversity of adversarial examples. Secondly, we generate adversarial text while simultaneously distancing it from the last intersection region along the optimization path. To investigate the impact of each improvement on the effectiveness of our method, we conduct ablation studies on the ITR task and employ transferability from ALBEF to the other three VLP models as the evaluation metric.

In the ablation study, we systematically eliminate each enhancement from our approach and compare their transferability with both the baseline SGA and our final method, which incorporates a combination of two improvements. The results are depicted in Figure 4. It is evident that when attacking models with the same architecture versus different architectures, the impacts of two distinct improvements are also different. When attacking TCL with the same structure as ALBEF, under **Setting1**, our method exhibits a greater decrease in transferability, indicating that diversification along the intersection region is

Table 4: Cross-Task Transferability. We utilize ALBEF to generate multimodal adversarial examples for attacking both Visual Grounding(VG) on the RefCOCO+ dataset and Image Captioning(IC) on the MSCOCO dataset. The baseline represents the performance of each task without any attack, where a lower value indicates better effectiveness of the adversarial attack for both tasks.

Attack	ITR → VG			ITR → IC				
	Val	TestA	TestB	B@4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline	58.46	65.89	46.25	39.7	31.0	60.0	133.3	23.8
SGA	50.56	57.42	40.66	28.0	24.6	51.2	91.4	17.7
Ours	**49.70**	**56.32**	**40.54**	**27.2**	**24.2**	**50.7**	**88.3**	**17.2**

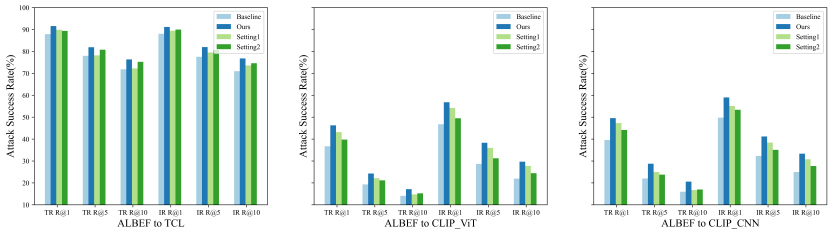

Fig. 4: Ablation Study: Attack Success Rate(%) on other three target models. The baseline is SGA. **Setting 1** removes diversification along the intersection region of adversarial trajectory. **Setting 2** removes the text deviating from the last intersection region along the optimization path.

most crucial in our approach under this scenario. However, under **Setting2**, the decrease in transferability is more pronounced when attacking the aligned VLP model, indicating that at this point, adversarial texts deviating from the last intersection region play a larger role in the effectiveness of our method.

5 Conclusion

In this paper, we conduct a systematic evaluation of existing multimodal attacks regarding transferability. We found that these methods predominantly prioritize diversity around adversarial examples (AEs) during the optimization process, potentially leading to overfitting to the victim model and hindering transferability. To address this issue, we propose diversification along the intersection region of adversarial trajectory to broaden diversity not only around AEs but also towards clean inputs. Moreover, we pioneer an exploration into extending adversarial texts deviating from the intersection region. Through extensive experiments, we demonstrate the effectiveness of our method in enhancing transferability across VLP models and V+L tasks. This work could act as a catalyst for more profound research on the transferability of multimodal AEs, alongside fortifying the adversarial robustness of VLP models.

Acknowledgments

This research is supported by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), Career Development Fund (CDF) of Agency for Science, Technology and Research (A*STAR) (No.: C233312028), and National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04).

References

1. Abdelfattah, R., Guo, Q., Li, X., Wang, X., Wang, S.: Cdul: Clip-driven unsupervised learning for multi-label image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1348–1357 (2023)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 382–398. Springer (2016)
3. Bai, Y., Zeng, Y., Jiang, Y., Wang, Y., Xia, S.T., Guo, W.: Improving query efficiency of black-box adversarial attack. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 101–116. Springer (2020)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
5. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12655–12663 (2020)
6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
7. Cheng, H., Xiao, E., Cao, J., Yang, L., Xu, K., Gu, J., Xu, R.: Typography leads semantic diversifying: Amplifying adversarial transferability across multimodal large language models. arXiv preprint arXiv:2405.20090 (2024)
8. Cheng, M., Sun, Y., Wang, L., Zhu, X., Yao, K., Chen, J., Song, G., Han, J., Liu, J., Ding, E., et al.: Vista: vision and scene text aggregation for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5184–5193 (2022)
9. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321 (2019)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. arXiv preprint arXiv:2010.11929 (2010)

11. Dou, Z.Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., et al.: An empirical study of training end-to-end vision-and-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18166–18176 (2022)
12. Gao, K., Bai, Y., Bai, J., Yang, Y., Xia, S.T.: Adversarial robustness for visual grounding of multimodal large language models. arXiv preprint arXiv:2405.09981 (2024)
13. Gu, J., Jia, X., de Jorge, P., Yu, W., Liu, X., Ma, A., Xun, Y., Hu, A., Khakzar, A., Li, Z., et al.: A survey on transferability of adversarial examples across deep neural networks. arXiv preprint arXiv:2310.17626 (2023)
14. Guo, Q., Cheng, Z., Juefei-Xu, F., Ma, L., Xie, X., Liu, Y., Zhao, J.: Learning to adversarially blur visual object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10839–10848 (2021)
15. Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Wang, J., Yu, B., Feng, W., Liu, Y.: Watch out! motion is blurring the vision of your deep neural networks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 975–985 (2020)
16. Han, D., Jia, X., Bai, Y., Gu, J., Liu, Y., Cao, X.: Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. arXiv preprint arXiv:2312.04403 (2023)
17. He, B., Jia, X., Liang, S., Lou, T., Liu, Y., Cao, X.: Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. arXiv preprint arXiv:2312.04913 (2023)
18. He, B., Liu, J., Li, Y., Liang, S., Li, J., Jia, X., Cao, X.: Generating transferable 3d adversarial point cloud via random perturbation factorization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 764–772 (2023)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
20. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling up vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17980–17989 (June 2022)
21. Huang, Y., Guo, Q., Juefei-Xu, F., Ma, L., Miao, W., Liu, Y., Pu, G.: Advfilter: predictive perturbation-aware filtering against adversarial attack via multi-domain learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 395–403 (2021)
22. Huang, Y., Juefei-Xu, F., Guo, Q., Zhang, J., Wu, Y., Hu, M., Li, T., Pu, G., Liu, Y.: Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 21169–21178 (2024)
23. Huang, Y., Sun, L., Guo, Q., Juefei-Xu, F., Zhu, J., Feng, J., Liu, Y., Pu, G.: Ala: Naturalness-aware adversarial lightness attack. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2418–2426 (2023)
24. Huang, Y., Sun, L., Guo, Q., Juefei-Xu, F., Zhu, J., Feng, J., Liu, Y., Pu, G.: Ala: Naturalness-aware adversarial lightness attack. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 2418–2426 (2023)
25. Jia, X., Wei, X., Cao, X., Han, X.: Adv-watermark: A novel watermark perturbation for adversarial examples. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1579–1587 (2020)

26. Jia, X., Zhang, Y., Wei, X., Wu, B., Ma, K., Wang, J., Cao Sr, X.: Improving fast adversarial training with prior-guided knowledge. arXiv preprint arXiv:2304.00202 (2023)
27. Khan, Z., Fu, Y.: Exploiting bert for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3034–3042 (2021)
28. Lei, C., Luo, S., Liu, Y., He, W., Wang, J., Wang, G., Tang, H., Miao, C., Li, H.: Understanding chinese video and language via contrastive multimodal pre-training. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2567–2576 (2021)
29. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
30. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
31. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: Bert-attack: Adversarial attack against bert using bert (2020)
32. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020)
33. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
34. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. arXiv preprint arXiv:1908.06281 (2019)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
36. Lu, D., Wang, Z., Wang, T., Guan, W., Gao, H., Zheng, F.: Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models (2023)
37. Luo, H., Gu, J., Liu, F., Torr, P.: An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. arXiv preprint arXiv:2403.09766 (2024)
38. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
39. Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Shahbaz Khan, F., Yang, M.H.: Intriguing properties of vision transformers. Advances in Neural Information Processing Systems **34**, 23296–23308 (2021)
40. Park, J., Miller, P., McLaughlin, N.: Hard-label based small query black-box adversarial attack. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3986–3995 (2024)
41. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
44. Shi, L., Shuang, K., Geng, S., Gao, P., Fu, Z., de Melo, G., Chen, Y., Su, S.: Dense contrastive visual-linguistic pretraining. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 5203–5212 (2021)
45. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
46. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
47. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
48. Wang, T., Ge, Y., Zheng, F., Cheng, R., Shan, Y., Qie, X., Luo, P.: Accelerating vision-language pretraining with free language modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23161–23170 (2023)
49. Wang, T., Jiang, W., Lu, Z., Zheng, F., Cheng, R., Yin, C., Luo, P.: Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In: International Conference on Machine Learning. pp. 22680–22690. PMLR (2022)
50. Wang, X., He, X., Wang, J., He, K.: Admix: Enhancing the transferability of adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16158–16167 (2021)
51. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5764–5773 (2019)
52. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2730–2739 (2019)
53. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)
54. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1307–1315 (2018)
55. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)
56. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 558–567 (2021)

57. Zhang, J., Yi, Q., Sang, J.: Towards adversarial attack on vision-language pre-training models. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5005–5013 (2022)
58. Zhang, J., Huang, J.t., Wang, W., Li, Y., Wu, W., Wang, X., Su, Y., Lyu, M.R.: Improving the transferability of adversarial samples by path-augmented method. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8173–8182 (2023)
59. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2021)
60. Zhang, Q., Lei, Z., Zhang, Z., Li, S.Z.: Context-aware attention network for image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3536–3545 (2020)